A Chi-square Statistic for Testing the Equality of Distracters' Plausibility in Multiple-Choice Test Items

Sherwin E. Balbuena

balbuenasherwine@debesmscat.edu.ph

Dr Emilio B Espinosa Sr Memorial State College of Agriculture and Technology

Research Article

Keywords: distracter analysis, chi-square test, item analysis, plausibility

Posted Date: June 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4441034/v1

Additional Declarations: No competing interests reported.

Abstract

This study introduces a new chi-square test statistic for testing the equality of response frequencies among distracters in multiple-choice tests. The formula uses the information from the number of correct answers and wrong answers, which becomes the basis of calculating the expected values of response frequencies per distracter. The method was applied to a statistics test response data and found to effectively detect unequally plausible distracters. Furthermore, the statistic had a quadratic relationship with item difficulty, indicating that at a certain range of plausibility values, there is an optimal item difficulty.

Introduction

Multiple-choice (MC) items are the most commonly used test type in schools due to their efficiency, objectivity, and ease of scoring [1][2]. These tests allow educators to assess a broad range of knowledge and skills quickly, making them ideal for large classrooms where individualized assessment can be time-consuming and impractical. Additionally, MC questions minimize the potential for scorer bias, providing a consistent and fair measure of student performance [3]. Their flexibility in testing various cognitive levels—from basic recall to application and analysis—further enhances their utility in diverse educational settings [4] [5].

Item analysis is a process of evaluating the quality of MC items. It involves statistical measures that provide information about their psychometric properties. The key item properties that are assessed include difficulty (DIFF), discrimination (DISC), and distracter efficiency (DE) [6] [7]. DIFF refers to the proportion of test-takers who answered an item correctly. The DISC indicates how well an item differentiates between high performers and low performers on the overall test, with higher DISC values indicating that more able examinees answered the items correctly. DE evaluates the functionality of incorrect options (distracters), ensuring that they plausibly attract those who do not know the correct answer while not misleading those who do. Together, these properties help refine multiple-choice items, enhancing their reliability and validity in measuring learners' knowledge and skills.

Distracter efficiency (DE) is an important metric in the research on multiple-choice item quality, focusing on the performance of incorrect answer options. A distracter is considered functional if it is selected by at least 5% of examinees, indicating that it is effective in attracting those who do not know the correct answer. This threshold ensures that each distracter contributes to the item's overall discriminative power [8] [9]. Some studies confirmed its strong relationship with DIFF and DISC [10] [8] [11], while others did not find a correlation with these item quality metrics [12]. However, this distracter-level parameter is estimated individually through the computation of the proportion of examinees who chose the distracter. There is a distinct lack of studies analyzing the collective efficiency of multiple distracters. Furthermore, there are no established methods to evaluate whether all distracters are equally plausible, although this property is recommended. This lack of equal plausibility metrics means that while individual distracters might meet the > 5% criterion, the quality of the corresponding item could still be compromised if the distracters do not function well together to attract less-able examinees.

This study aimed to develop a new method of assessing distracter plausibility relative to the response frequencies for the correct option and the other distracters. The new method will be tested for correlations with other item parameters, such as difficulty and discrimination, to determine whether they independently or collectively measure item quality.

Methods

Research Design

This study employs an exploratory research design, which focuses on the development of a new statistical method for item analysis. It also aims to explore whether the statistical measure can be used as a supplementary assessment tool to assess the quality of multiple-choice test items. Correlations of the new statistic with other item analysis metrics will be conducted to assess the former's validity and possible complementary role in item quality checks.

Study Context

The method used in this study was applied to the analysis of items in a statistical test, which measures the student's ability to choose appropriate parametric and nonparametric techniques given the data characteristics or assumptions met. This test was administered in a graduate-level statistics course at one state college in the Bicol region, the Philippines, covering the school years 2021–2022 to 2022–2023. The 6-item test comprises only a portion of the final exams after excluding extremely easy and extremely difficult items and other non-MC items.

Calculation of Item Parameters

The calculations of DIFF, DISC, and the new statistic, referred to in this study as *M*, were computed using different approaches. DIFF and DISC were estimated using Rasch model analysis in the eRm package in R [13] [14]. The *M* statistic is given by the formula below, which is based on the calculation of the chi-square statistic using an expected value obtained by dividing the number of incorrect responses to an item by the number of distracters.

$$M = \sum_{j=1}^{d} z_j = \sum_{j=1}^{d} \frac{(n_j - e_j)^2}{e_j}$$

where

$n_j =$ observed frequency of responses for distracter *j*

$e_j = \frac{w_i}{d} = \frac{N_i - c_i}{d}$ = expected value for distracter *j*

w_i = number of wrong responses for item *i*

N_i = number of test takers for item *i*

c_i = number of correct answers for item *i*

The degrees of freedom are *d* or the number of distracters.

Data Analysis

The obtained values of DIFF, DISC, and M were correlated using Pearson product-moment correlation to determine whether a linear relationship existed. Further modeling using polynomial regression was conducted to obtain a better fit of the model to the data. The analyses were all performed in R and its IDE RStudio. All levels of significance were set at 5%.

Results and Discussion

Using the frequency of correct responses, the M statistic can be derived by dividing the frequency of incorrect responses by the number of distracters. In Table 1, sample computations of the expected values are presented. For a number of distracters d = 3, a number of examinees N = 100, and a number of correct responses c = 70, for example, the expected frequency per distracter is 10.0. Given a certain observed frequency, its distance from the corresponding expected value can be computed, which may be a negative or positive distance. Squaring these differences and dividing by the expected value results in a chi-square with 1 degree of freedom. Summing these ratios of squared differences and expectations across the number of distracters *d* results in a chi-square test with *d* degrees of freedom.

Table 1

Sample expected values of response frequencies for items with d = 3 and N = 100

No. of correct responses (c)	**Expected value of number of responses for a distracter** **[(N-c)/3]**	**No. of correct responses (c)**	**Expected value of number of responses for a distracter** **[(N-c)/3]**
70	10.0000	78	7.3333
71	9.6667	79	7.0000
72	9.3333	80	6.6667
73	9.0000	81	6.3333
74	8.6667	82	6.0000
75	8.3333	83	5.6667
76	8.0000	84	5.3333
77	7.6667	85	5.0000

An example of whether the hypothetical item distracters are equally plausible based on their frequencies is provided in Table 2. For instance, with 25 correct responses (c), the number of incorrect responses totals N – c, which is 100–25 = 75. If Distracters 1, 2, and 3 receive 35, 15, and 25 responses, respectively, and the expected value for each distracter is 8.33, their respective z values are calculated as 4, 4, and 0, summing to 8. The corresponding p-value for this statistic with df = 3 is 0.046, indicating that the observed frequencies significantly differ from the expected frequencies. Therefore, the distracters are not equally plausible.

Table 2

An illustrative example of applying the statistic (d = 3, N = 100)

c	n1	n2	n3	z1	z2	z3	chisq	p value	Equally plausible?
25	35	15	25	4.0000	4.0000	0.0000	8.0000	0.046	No
25	46	18	11	27.3282	0.6205	5.2513	33.2000	0.000	No
25	51	7	17	41.2552	9.6302	0.8802	51.7656	0.000	No
26	20	20	34	0.8829	0.8829	3.5315	5.2973	0.151	Yes
28	4	50	18	16.6667	28.1667	1.5000	46.3333	0.000	No
32	15	15	38	2.5931	2.5931	10.3725	15.5588	0.001	No
34	20	23	23	0.1818	0.0455	0.0455	0.2727	0.965	Yes

Application to a Dataset

We analyzed a dataset consisting of test responses from a graduate-level statistics course with 198 participants. For six items, we recorded the frequency of responses for each option, marking the frequency of correct responses with an asterisk "*". Using the *M* statistic, we computed the expected values, which are displayed in Table 3. For instance, in Item 1, 104 out of 198 examinees selected the correct answer, while 94 chose incorrect answers (distracters). Based on these data, the expected value was approximately 31.33.

Table 3

Frequencies (percentages) of response to all options for the 6 test items and the corresponding expected frequency per distracter

Item	Options				N-c	Expected value
	A	B	C	D		
1	104*(53%)	37(19%)	41(21%)	16(8%)	94	31.33
2	37(19%)	103*(52%)	19(10%)	39(20%)	95	31.67
3	20(10%)	16(8%)	142*(72%)	20(10%)	56	18.67
4	35(18%)	40(20%)	20(10%)	103*(52%)	95	31.67
5	11(6%)	11(6%)	135*(68%)	41(21%)	63	21.00
6	11(6%)	6(3%)	36(18%)	145*(73%)	53	17.67
* frequency of correct responses (c)						

The DISC and DIFF parameters were estimated using the dichotomous Rasch model to evaluate the psychometric properties of the items. As shown in Table 4, two items (Items 2 and 5) had negative discrimination values, indicating poor quality. There were three easy items (Items 3, 5, and 6 with negative logits) and three difficult items (Items 1, 2, and 4 with positive logits). Further analysis using M revealed that three items had equally plausible distracters (Items 2, 3, 4), while three items had distracters of unequal plausibility (Items 1, 5, 6). Item 5 was flagged as poor quality due to both its nondiscriminative nature and unequally plausible distracters. Overall, only Items 3 and 4 demonstrated good quality based on the assessed properties.

In this new approach, the detection of implausible distracters is different from the traditional approach popularized by Haladyna and Downing (1993) [8]. Although most frequencies exceeded the > 5% criterion for functional distracters, except for Distracter B of Item 6, which had a 3% response, the items were still flagged for being collectively ineffective in attracting less-able test takers. Hence, the new method can complement the existing methodologies in distracter analysis to identify items with dysfunctional distracters for further investigation.

Table 4

Rasch-based Item DISC and DIFF estimates and results of equality of distracters' plausibility tests

Item	DISC	DIFF	M	p-value	Equally plausible?
1	0.244	0.519	11.51064	0.009	No
2	-0.062	0.546	7.663158	0.053	Yes
3	0.364	-0.577	0.571429	0.903	Yes
4	0.150	0.546	6.842105	0.077	Yes
5	-0.103	-0.358	28.57143	0.000	No
6	0.475	-0.676	29.24528	0.000	No

The flagged items considered for revision are shown in Table 5. The contents of the 3 problematic items below are given with the recommended revisions at the distracter level. Item 1 was found to have unequally plausible distracters, as shown in Table 4; hence, distracter D with the least frequency was revised from "Wilcoxon signed-rank test" to "Welch t-test". The original option D was not attractive, possibly due to its association with paired data analysis. Replacing this with the "Welch t test" may be more effective since the tool is used as an alternative when the assumption for homogeneity of variances is violated.

Table 5

Item content and options of the three items flagged for unequal plausibility and recommended revisions with justifications

Item Content and Options (with recommended revisions)	Reason/s for revision
Item 1. The following are the characteristics of data: (1) dependent variable is measured at interval/ratio level; (2) There are two independent categories for the nominal independent variable; (3) The distribution in each group is normal; (4) The variances of the two groups are equal; (5) There are no observed outliers. What statistical tool is the most appropriate to compare the two groups? A. Independent groups t test* B. Mann−Whitney U test C. Paired t test D. Wilcoxon signed-rank test (Replace with Welch t test)	Unequal plausibility of distracters. Distracter D had the lowest frequency. The replacement is assumed to distract more effectively because the tool is used alternatively when t test Assumption (4) is not met.
Item 5. The following are the characteristics of data: (1) Variables X and Y are measured at interval/ratio level; (2) X and Y are paired; (3) The distribution of the paired data are bivariate normal; (4) There are no observed outliers; (5) There is a linear relationship between X and Y. What statistical tool is the most appropriate to test the hypothesis that there is no linear correlation between X and Y? A. Analysis of variance (Replace with Chi-square test of independence) B. Paired t test (Replace with Point-biserial correlation) C. Pearson product-moment correlation* D. Spearman rank correlation	Unequal plausibility of distracters. Distracters A and B had the lowest frequencies maybe because the contents are tests of comparison. The replacements are assumed to distract more effectively because the tools are used alternatively to test relationships between variables.
Item 6. The following are the characteristics of data: (1) Variables X and Y are measured at interval/ratio level; (2) X and Y are paired; (3) The distribution of the paired data are not normal; (4) There are observed outliers; (5) There is a monotonic relationship between X and Y. What statistical tool is the most appropriate to test the hypothesis that there is no correlation between X and Y? A. Analysis of variance (Replace with Chi-square test of independence) B. Paired t test (Replace with Point-biserial correlation) C. Pearson product-moment correlation D. Spearman rank correlation*	Unequal plausibility of distracters. Distracters A and B had the lowest frequencies maybe because the contents are tests of comparison. The replacements are assumed to distract more effectively because the tools are used alternatively to test relationships between variables.

In items involving the assumptions of tests of relationships between variables, Items 5 and 6 had unequally plausible distracters. Two of the three distracters were comparison tests (e.g., ANOVA and paired t tests); therefore, they were less appealing or less effective as distracters in a collective manner.

This is likely because these types of tests may not be as relevant or plausible within the context of the question (e.g., assumptions of correlation test), making them less likely to be chosen by examinees who do not know the correct answer. Consequently, these distracters fail to effectively challenge test takers and are not as efficient at diverting them from the correct answer. Replacing these with other tests of relationships (e.g., chi-square test of independence and point-biserial correlation) may address this unequal plausibility.

Correlation of M with DIFF and DISC

To investigate potential relationships between DIFF and M, as well as between item DISC and M, we conducted correlation and regression analyses. The results indicated a moderate negative linear correlation between M and DIFF (r = -0.458), although this relationship was not statistically significant (p > 0.05). This suggests that as item difficulty increases, the plausibility of the distracters tends to decrease, but the relationship is not strong enough to be conclusive. Additionally, no significant correlation was found between M and DISC (r = -0.037, p > 0.05), indicating that the discriminative power of an item is not related to the plausibility of its distracters. Finally, there was a nonsignificant negative correlation between DIFF and DISC (r = -0.465, p > 0.05), implying that while there may be a tendency for more difficult items to be less discriminative, this trend is not statistically significant. Overall, the analyses suggest linear independence among these metrics.

Despite the lack of correlations, we noted some polynomial trends in the relationships between *M* and DIFF. The scatterplot in Fig. 1 shows a rather curvilinear trend such that when the value of the M statistic changes, the value of DIFF follows an inverted U pattern with a maximum a value near M = 10. Therefore, we conducted a polynomial regression to determine if a polynomial function fits the empirical data.

Table 6

Results of polynomial regression showing significant coefficients for linear and quadratic trends

Coefficients:	Estimate	Std. error	t value	p value
(Intercept)	-0.610253	0.187745	-3.250	0.04747 *
M	0.189268	0.033788	5.602	0.01124 *
I(M^2)	-0.006447	0.001007	-6.400	0.00773 **

Residual standard error: 0.1791 for 3 degrees of freedom

Multiple R-squared: 0.9461, Adjusted R-squared: 0.9101

F-statistic: 26.31 on 2 and 3 DF, p value: 0.01253

In Table 6, polynomial regression showed a multiple R-squared of 0.9461, which indicates that approximately 94.61% of the variance in DIFF is explained by the model, showing a high degree of fit. The adjusted R-squared equals 0.9101 after adjusting the R-squared value for the number of predictors in the model, still indicating a strong fit. The overall model significance test suggested that the model significantly predicted DIFF [F(2,3) = 26.31, p = 0.01253].

The polynomial regression results reveal a significant quadratic relationship between DIFF and M. The significant negative coefficient for M^2 suggests a parabolic curve where DIFF initially increases with M but starts to decrease as M continues to increase. The high R-squared and adjusted R-squared values indicate that the model explains a large portion of the variability in DIFF. Given the significant p values for all the coefficients, we can infer that the equal plausibility metric (M) has a meaningful and complex impact on item difficulty (DIFF). This model can be useful for understanding how changes in M affect DIFF and can inform the design and evaluation of test items to achieve desired levels of difficulty.

Conclusion and Recommendations

This study introduced a chi-square statistic, M, designed to detect significant deviations from the expected frequencies of distracters, known as the equal plausibility of distracters. This novel item analysis metric fills a gap by evaluating the collective functionality of distracters and serves as a basis for identifying items with dysfunctional distracters. The statistic was empirically tested using response data from a statistics test and was found to effectively detect items with implausible distracters. Furthermore, the new metric showed a quadratic relationship with item difficulty, suggesting an optimal difficulty level within a specific range of M values.

Several limitations were noted in this study. First, the dataset included only 6 items, which is a very small sample size for item analysis, potentially affecting the observed relationships. Second, DIFF and DISC estimates were obtained using the Rasch model, which differs from classical test theory estimates. No assumption tests were conducted to confirm whether the items met the Rasch model's expectations, potentially invalidating the derived estimates. Future research is encouraged to verify these results and address the limitations identified in this study.

Declarations

Funding

This study did not receive funding from any granting agencies.

Human Ethics and Consent to Participate

This study utilized a test response data set to evaluate the applicability of the developed statistical method. No personal or identifiable information was used, ensuring that ethical standards were maintained. As such, formal consent from participants was not required for this analysis.

Consent for Publication

Not applicable.

Competing Interest

The author declares no competing interests.

Data Availability Statement

The data used in this study will be available upon request.

Author Contribution

The author solely conceptualized the study, conducted the data analysis, and authored the manuscript. All aspects of the research, from the initial idea through to the final write-up, were independently carried out by the author.

References

1. Abdulghani, H. M., Irshad, M., Haque, S., Ahmad, T., Sattar, K., & Khalil, M. S. (2017). Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. *PloS One, 12*(10), e0185895.
2. Wood, E., Klausz, N., & MacNeil, S. (2022). Examining the influence of multiple-choice test formats on student performance. *Innovative Higher Education, 47*, 515–531. https://doi.org/10.1007/s10755-021-09581-7
3. Xu, X., Kauer, S., & Tupy, S. (2016). Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarship of Teaching and Learning in Psychology*, *2*(2), 147.
4. Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple-choice questions? Research paper. *BMC Medical Education, 7*, 1-7.
5. Mitra, A. K. (2022). The Art of Designing a Quality Multiple Choice Question in Chemistry. *Resonance*, *27*(6), 1017-1031.
6. Elgadal, A. H., & Mariod, A. A. (2021). Item analysis of multiple-choice questions (MCQs): assessment tool for quality assurance measures. *Sudan Journal of Medical Sciences, 16*(3), 334-346.
7. DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4.
8. Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item?. *Educational and Psychological Measurement, 53*(4), 999-1010.
9. Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distracters for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116.

10. Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., ... & Adam, M. I. E. (2024). Item analysis: the impact of distracter efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education, 24*(1), 445.
11. Testa, S., Toscano, A., & Rosato, R. (2018). Distracter efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology, 9*, 357601.
12. Puthiaparampil, T., & Rahman, M. (2021). How important is distracter efficiency for grading Best Answer Questions?. *BMC Medical Education, 21*, 1-6.
13. Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*, 1-20.
14. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Figures

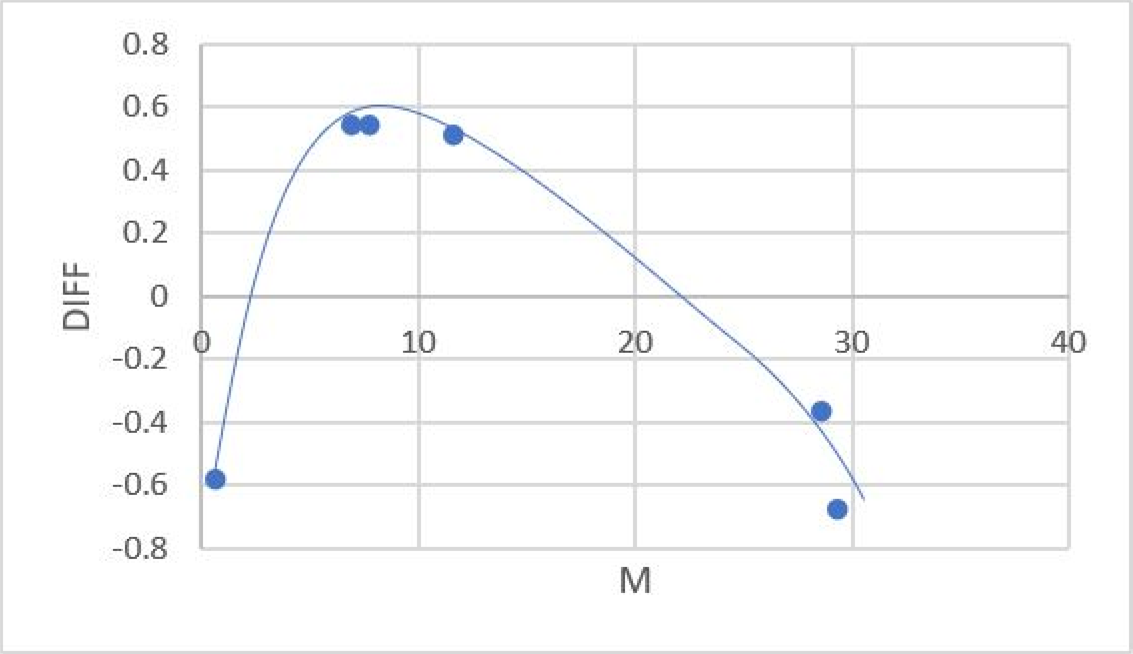

Figure 1

Scatterplot of M and DIFF showing the curvilinear relationship